SPS 2467 Statistical Model Building

Generalized Linear Models

Dr. Mutua Kilai

Kirinyaga University

Review and Introduction

Let $y_1, ..., y_n$ denote n independent observations on a response.

Treat y_i as a realization of a random variable Y_i

In the general linear model we assume that

$$Y_i \sim N(\mu_i, \sigma^2)$$

And we further assume that the expected value μ_i is a linear function

$$\mu_i = X_i'\beta$$

The generalized linear model generalizes both the random and systematic component.

Components of Generalized Linear Models

All generalized linear models have three components:

- Random component
- Systematic component
- Link function

Random Component

The random component of a GLM identifies the response variable Y and selects a probability distribution for it.

Denote the observations on Y by $(Y_1, Y_2, ..., Y_n)$. Standard GLMs treat $Y_1, Y_2, ..., Y_n$ as independent.

If the observations on Y are binary then we assume a *binomial distribution* for Y

In some applications, each observation is a count. Then we have *Poisson or Negative Binomial*

If each observation is continuous, we might assume a normal distribution for Y.

Systematic Component

The systematic component of a GLM specifies the explanatory variables.

These enter linearly as predictors on the right-hand side of the model equation.

The systematic component specifies the variables that are the $\{x_j\}$ in the formula

$$\alpha + \beta_1 x_1 + ... + \beta_k x_k$$

Link Function

Denote the expected value of Y the mean of the probability distribution by $\mu = E(Y)$

The link function specifies a function $g(.)$ that relates μ to the linear predictors as

$$g(\mu) = \alpha + \beta_1 x_1 + ... + \beta_k x_k$$

The function $g(\mu)$ the link function connects the random and the systematic components.

The Exponential Family

We assume that observations come from a distribution in the exponential family with the following probability density function:

$$f(y_i; \theta_i, \phi) = exp\Big\{\frac{y_i\theta_i}{a(\phi)} + c(y_i, \phi)\Big\} \tag{1}$$

Here θ_i, ϕ are parameters and $a(.), b(.)$ and $c(.)$ are known functions.

The θ_i and ϕ are location and scale parameters respectively.

Normal Distribution

The normal distribution is given as:

$$f(y_i, \theta_i, \phi) = \frac{1}{\sqrt{2\pi\sigma}} exp\{-\frac{1}{2\sigma^2}(y_i - \mu)^2\}$$

Which can be expressed as:

$$f(y_i, \theta_i, \phi) = \exp\Big[-\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(y_i^2 - 2y_i\mu + \mu^2)\Big]$$

We can re-factor and have:

$$f(y_i, \theta_i, \phi) = \Big(\frac{2\mu y_i - \mu^2}{2\sigma^2}\Big) - \frac{1}{2}\Big(\frac{y_i^2}{\sigma^2} + \log(2\pi\sigma^2)\Big)$$

$\theta_i = \mu, \phi = \sigma^2, a_i(\phi) = \phi, b(\theta_i) = \frac{\theta_i^2}{2}, c(y_i, \phi) = \frac{1}{2}\Big(\frac{y_i^2}{\sigma^2} + \log(2\pi\sigma^2)\Big)$

The mean is given as $E(y_i) = b'(\theta_i)$

The variance $Var(y_i) = b''(\theta_i)a(\phi)$

Exercises

Exercise 1

The PMF of the Poisson distribution is given as:

$$f(y|\mu) = \frac{e^{-\mu}\mu^y}{y!}$$

Show that the Poisson Distribution can be expressed as a member of exponential family and derive the mean and variance.

Exercise 2

The PMF of the Binomial distribution is given as:

$$f(y|n,p) = \binom{n}{y} p^y (1-p)^{n-y}$$

Show that the binomial Distribution can be expressed as a member of exponential family and derive the mean and variance.

Exercise 3

The PMF of the Negative Binomial distribution is given as:

$$f(y|r,p) = \binom{r+y-1}{y} p^r (1-p)^y$$

Show that the negative binomial Distribution can be expressed as a member of exponential family and derive the mean and variance.

Maximum Likelihood Estimation of GLM

Unlike for the general linear model, there is no closed form expression for the MLE of β in general for GLMs.

However all the GLMs can be fit using the same algorithm a form of iteratively re-weighted least squares

Given an initial value for $\hat{\beta}$ calculate the estimated linear predictor $\hat{\eta_i} = x_i'\beta$ and use that to obtain the fitted values $\hat{\mu_i} = g^{-1}(\hat{\eta_i})$. Calculate the adjusted dependent variable

$$z_i = \hat{\eta_i} + (y_i - \hat{\mu_i})(\frac{d\eta_i}{d\mu_i})_0$$

Calculate the iterative weights

$$W_i^{-1} = (\frac{d\eta_i}{d\mu_i})V_i$$

where V_i is the variance function evaluated at $\hat{\mu_i}$

Regress z_i on x_i with weight W_i to give the new estimate of β

Logistic Regression

In logistic problems we are modeling binary data. The usual coding is that

$$Y \in \{1 = "Success \ \ or \ \ 0 = "Failure\}$$

The *Binomial* distribution is a good way to represent this kind of data.

The systematic component in our logistic regression model will be the binomial distribution.

We show that the binomial distribution belongs to the exponential family of distributions

$$\begin{aligned} f(y;\theta,\phi) &= \binom{n}{y}\pi^y(1-\pi)^{n-y} \\ &= \exp\left[y\log(\frac{\pi}{1-\pi}) + n\log(1-\pi) + \log\binom{n}{y}\right] \end{aligned} \tag{2}$$

Here

$$\theta = \log(\frac{\pi}{1-\pi})$$

$$b(\theta) = -\log(1-\pi) = \log(1+\exp(\theta))$$

$$\mu = b'(\theta) = \frac{\partial}{\partial\theta}\log[1+\exp(\theta)] = \frac{\theta}{1+\exp(\theta)} = \pi$$

$$g(\mu) = \log[\frac{\pi}{1-\pi}] = \theta$$

You can easily show that

$$E[Y_i] = \mu_i = n_i\pi_i$$

and

$$Var(Y_i) = \sigma_i^2 = n_i\pi_i(1-\pi_i)$$

In logistic regression the outcome is binary example

- Alive or dead
- Pass or fail
- Pay or Default

Logit Transformation

We would like to have the probabilities π_i depend on a vector of observed covariates X_i

The idea is to let π_i be a linear function of the covariates say

$$\pi_i = X_i'\beta$$

where β is a vector of regression coefficients.

We transform the probability π_i to have the odds defined as:

$$odds_i = \frac{\pi_i}{1-\pi_i}$$

Taking the natural logarithm of the odds that is *logit* or log-odds we have:

$$\eta_i = logit\pi_i = \log\frac{\pi_i}{1-\pi_i}$$

Solving for π_i we have:

$$\pi_i = logit^{-1}(\eta_i) = \frac{e^{\eta_i}}{1+e^{\eta_i}}$$

We are now in a position to define the logistic regression model by assuming that the *logit* of the probability π_i rather than the probability itself follows a ,linear model.

Logistic Regression Model

Suppose that we have k independent observations $y_1, ..., y_k$ and that the $i-th$ observation can be treated as a realization of the random variable Y_i.

We assume that Y_i has a binomial distribution

$$Y_i \sim B(n_i, \pi_i)$$

The above equation defines the stochastic structure of the model.

Suppose further that the *logit* of the underlying probability π_i is a linear function of the predictors

$$logit(\pi_i) = x_i'\beta$$

Where x_i is a vector of covariates and β is a vector of regression coefficients. This defines the systematic structure of the model.

The models defined above is a generalized linear model with binomial response and link logit.

The interpretation of β_j represents the change in the *logit* of the probability associated with a unit change in the $j-th$ predictor holding all other predictors constant.

Exponetiating equation above we find the odds for the $i-th$ unit given by

$$\frac{\pi_i}{1-\pi_i} = \exp\{x_i'\beta\}$$

This expression defines a multiplicative model for the odds.

Exponentiating we get $\exp\{x_i'\beta\}$ times $\exp\{\beta_j\}$.

The exponentiated $\exp\{\beta_j\}$ represents the *odds ratio*

Solving for the probability π_i in the logit model gives the more complicated model

$$\pi_i = \frac{\exp\{x_i'\beta\}}{1+\exp\{x_i'\beta\}}$$

Estimation and Hypothesis Testing

Maximum Likelihood Estimation

The likelihood function for n independent binomial observations is a product of densities.

Taking logs, we find that the log-likelihood function

$$\log L(\beta) = \sum\{y_i \log(\pi_i) + (n_i - y_i)\log(1-\pi_i)\}$$

where π_i depends on the covariates x_i and a vector of p parameters β through the logit transformation.

The working dependent variable z_i which has elements

$$z_i = \hat{\eta}_i + \frac{y_i - \hat{\mu}_i}{\hat{\mu}_i(\eta_i - \hat{\mu}_i)}n_i$$

Where n_i are the binomial denominators. We then regress z on the covariates calculating the weighted least squares estimate

$$\hat{\beta} = (X'WX)^{-1}X'Wz$$

Where W is a diagonal matrix of weights with entries

$$w_{ii} = \hat{\mu}_i(n_i - \hat{\mu}_i)/n_i$$

The variance is given by:

$$var(\hat{\beta}) = (X'WX)^{-1}$$

Goodness of Fit Statistic

Suppose we have just fitted a model and want to assess how well it fits the data.

A measure of discrepancy between observed and fitted values is the deviance statistic, which is given by

$$D = 2\sum\{y_i \log(\frac{y_i}{\hat{\mu}_i}) + (n_i - y_i)\log\Big(\frac{n_i - y_i}{n_i - \hat{\mu}_i}\Big)\} \tag{3}$$

where y_i is the observed and $\hat{\mu}_i$ is the fitted value for the $i-th$ observation.

An alternative measure of goodness of fit is *Pearson chi-squared statistic* which for binomial data can be written as

$$\chi^2_P = \sum_i \frac{n_i(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i(n_i - \hat{\mu}_i)} \tag{4}$$

Tests of Hypothesis

As usual, we can calculate Wald tests based on the large-sample distribution of the m.l.e., which is approximately normal with mean β and variance-covariance matrix.

In particular we can test the hypothesis,

$$H_0 : \beta_j = 0$$

Concerning the significance of a single coefficient by calculating the ratio of the estimate to its standard error

$$z = \frac{\hat{\beta}_j}{\sqrt{\hat{Var}(\hat{\beta}_j)}}$$

This statistic has approximately a standard normal distribution in large samples.

The wald test can be use to calculate a confidence interval for β_j

The $100(1-\alpha)\%$ confidence that the true parameter lies in the interval with boundaries

$$\hat{\beta}_j \pm z_{1-\frac{\alpha}{2}} \sqrt{\hat{Var}(\hat{\beta}_j)}$$

Confidence intervals for effects in the logit scale can be translated into confidence intervals for odds ratios by exponentiating the boundaries.

Example 1

A researcher is interested in how variables, such as GRE (Graduate Record Exam scores), GPA (grade point average) and prestige of the undergraduate institution, effect admission into graduate school. The response variable, admit/don't admit, is a binary variable.

```
mydata <- read.csv("admit.csv")
knitr::kable(head(mydata))
```

admit	gre	gpa	rank
0	380	3.61	3
1	660	3.67	3
1	800	4.00	1
1	640	3.19	4
0	520	2.93	4
1	760	3.00	2

The code below estimates a logistic regression model using the glm (generalized linear model) function. First, we convert rank to a factor to indicate that rank should be treated as a categorical variable.

```
# convert rank to factor
mydata$rank <- factor(mydata$rank)

# fit the logistic regression model
mylogit <- glm(admit ~ gre + gpa + rank, data = mydata, family = "binomial")

# output a summary table neatly
```

```
library(gtsummary)

# output without odds ratio

tbl_regression(mylogit)
```

Characteristic	log(OR)	95% CI	p-value
gre	0.00	0.00, 0.00	0.038
gpa	0.80	0.16, 1.5	0.015
rank			
1	—	—	
2	-0.68	-1.3, -0.06	0.033
3	-1.3	-2.0, -0.67	<0.001
4	-1.6	-2.4, -0.75	<0.001

The logistic regression coefficients give the change in the log odds of the outcome for a one unit increase in the predictor variable.

Both gre and gpa are statistically significant, as are the three terms for rank.

- For every one unit change in gre, the log odds of admission (versus non-admission) increases by 0.002
- For a one unit increase in gpa, the log odds of being admitted to graduate school increases by 0.804
- The indicator variables for rank have a slightly different interpretation. For example, having attended an undergraduate institution with rank of 2, versus an institution with a rank of 1, changes the log odds of admission by -0.675.

We can test for an **overall effect** of rank using the **wald.test** function.

```
library(aod)
wald.test(b = coef(mylogit), Sigma = vcov(mylogit), Terms = 4:6)
```

```
## Wald test:
## ----------
##
## Chi-squared test:
## X2 = 20.9, df = 3, P(> X2) = 0.00011
```

The chi-squared test statistic of 20.9, with three degrees of freedom is associated with a p-value of 0.00011 indicating that the overall effect of rank is statistically significant.

The odds ratio with their respective CI is given as

```
# table with odds ratio
library(gtsummary)

tbl_regression(mylogit, exponentiate = TRUE)
```

Characteristic	OR	95% CI	p-value
gre	1.00	1.00, 1.00	0.038
gpa	2.23	1.17, 4.32	0.015
rank			
1	—	—	
2	0.51	0.27, 0.94	0.033

Characteristic	OR	95% CI	p-value
3	0.26	0.13, 0.51	<0.001
4	0.21	0.09, 0.47	<0.001

Now we can say that for a one unit increase in gpa, the odds of being admitted to graduate school (versus not being admitted) increase by a factor of 2.23.

The fitted model is given by:

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_{i2} + \beta_{4X_{i3}} + \beta_{5X_{i4}}$$

$$\log\left(\frac{\pi}{1-\pi}\right) = -3.98 + 0.002 X_1 + 0.80 X_2 - 0.68 X_3 - 1.3 X_4 - 1.6 X_5$$

We can predict a new variable log of the odds and have:

If the gpa score is 3.8, gre score is 4.0 and the rank of the student is 2

Then:

$$\log\left(\frac{\pi}{1-\pi}\right) = -3.98 + 0.002 \times 4 + 0.80 \times 3.8 - 0.68 \times 1 - 1.3 \times 0 - 1.6 \times 0$$

The odds is the exponentiate of the log-odds as follows:

$$\pi = \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\beta_3 X_{i2}+\beta_{4X_{i3}}+\beta_{5X_{i4}}}}{1+e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\beta_3 X_{i2}+\beta_{4X_{i3}}+\beta_{5X_{i4}}}}$$

Likelihood Ratio Test

The **likelihood ratio test** is used to test the null hypothesis that any subset of $\beta^{'s}$ is equal to zero.

The likelihood ratio test statistic is given as

$$\Lambda^* = -2(L(\hat{\beta^{(0)}}) - L(\hat{\beta}))$$

where $l(\hat{\beta})$ is the log-likelihood of the fitted model $l(\hat{\beta^{(0)}})$ is the log-likelihood of the reduced model specified by the null hypothesis evaluated at the maximum likelihood estimate of that reduced model.

The test statistic has a χ^2 distribution with $k-r$ degrees of freedom.

Statistical software often presents results for this test in terms of deviance, which is defined as -2 times log-likelihood.

We can compare the two models as:

- Fit one model without the rank variable
- Fit another model with the rank variable

```
# model 1

model1 <- glm(admit ~ gre + gpa, data = mydata, family = "binomial")
summary(model1)
```

```
##
## Call:
## glm(formula = admit ~ gre + gpa, family = "binomial", data = mydata)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.949378   1.075093  -4.604 4.15e-06 ***
## gre          0.002691   0.001057   2.544   0.0109 *
## gpa          0.754687   0.319586   2.361   0.0182 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 499.98  on 399  degrees of freedom
## Residual deviance: 480.34  on 397  degrees of freedom
## AIC: 486.34
##
## Number of Fisher Scoring iterations: 4
```

```
# Model 2

# convert rank to factor
mydata$rank <- factor(mydata$rank)

# fit the logistic regression model
model2 <- glm(admit ~ gre + gpa + rank, data = mydata, family = "binomial")
summary(model2)
```

```
##
## Call:
## glm(formula = admit ~ gre + gpa + rank, family = "binomial",
##     data = mydata)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.989979   1.139951  -3.500 0.000465 ***
## gre          0.002264   0.001094   2.070 0.038465 *
## gpa          0.804038   0.331819   2.423 0.015388 *
## rank2       -0.675443   0.316490  -2.134 0.032829 *
## rank3       -1.340204   0.345306  -3.881 0.000104 ***
## rank4       -1.551464   0.417832  -3.713 0.000205 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 499.98  on 399  degrees of freedom
## Residual deviance: 458.52  on 394  degrees of freedom
## AIC: 470.52
##
## Number of Fisher Scoring iterations: 4
```

The deviance statistic is 480.34 - 458.52 = 21.82. The $\chi^2_{1,0.05} = 3.84$ thus we conclude that the full model is better than the reduced model.